University of Southern Denmark
IMADA
Arthur Zimek
Jonatan Møller Gøttcke, Jonas Herskind Sejr

DM566/DM868/DM870: Data Mining and Machine Learning

Spring term 2019

Exercise 5: Distance Measures

Exercise 5-1 Distance functions

Distance functions can be classified into the following categories:

$d : S \times S \to \mathbb{R}_0^+$ $x, y, z \in S :$	reflexive $x = y \Rightarrow d(x, y) = 0$	symmetric $d(x, y) = d(y, x)$	strict $d(x, y) = 0 \Rightarrow x = y$	triangle inequality $d(x, z) \leq d(x, y) + d(y, z)$
Dissimilarity function	$\times$			
(Symmetric) Pre-metric	$\times$	$\times$		
Semi-metric, Ultra-metric	$\times$	$\times$	$\times$	
Pseudo-metric	$\times$	$\times$		$\times$
Metric	$\times$	$\times$	$\times$	$\times$

So if a distance measure satisfies $d : S \times S \to \mathbb{R}_0^+$ and $\forall x, y, z \in S$ it is reflexive, symmetric, and strict and it also satisfies the triangle inequality, then it is a metric.

As you can see, a pre-metric does not necessarily need to be *strictly* reflexive. Make sure you understand the difference between reflexivity and strictness!

Note: these terms as well as "distance function" are used inconsistently in the literature. In mathematics, "distance function" is commonly used synonymously with "metric". In a database and data mining context, strictness is often not relevant at all, and a "distance function" usually refers to a pseudo-metric, pre-metric, or even just to some dissimilarity function. Do not rely on Wikipedia, it uses multiple definitions within itself!

Decide for each of the following functions $d(\mathbb{R}^n, \mathbb{R}^n)$, whether they are a distance, and if so, which type.

(a) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)$

(b) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$

(c) $d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$

(d) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i = y_i \\ 0 & \text{iff} \quad x_i \neq y_i \end{cases}$

(e) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i \neq y_i \\ 0 & \text{iff} \quad x_i = y_i \end{cases}$

Exercise 5-2 Induced metric

Given a pseudo-metric d on the set A: $d : A \times A \to \mathbb{R}_0^+$.

Define the equivalence relation $\sim$ such that $x \sim y \Leftrightarrow d(x, y) = 0$.

Let $A^\sim$ be the set of equivalence classes of A w.r.t. $\sim$.

- Which properties has the distance function $d^\sim : A^\sim \times A^\sim \to \mathbb{R}_0^+$ with $d^\sim(x^\sim, y^\sim) := d(x, y)$?
- Given a database similar to this one:

r	x	y
1	0	1
2	1	1
3	0	1

r	x	y
4	1	1
5	2	2
6	3	3

Which properties does the following distance function have?

$$\text{euclid}_{xy}((r_1, x_1, y_1), (r_2, x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Explain which records are considered equivalent by this distance function, and discuss whether it is sensible in a database and data mining context to have pseudo-metric distance functions.

Hint: What could be the nature of attribute r in a database context?